An Overview on Variable Selection for Survival Analysis

Article · March 2005

DOI: 10.1142/9789812567765_0019

3 authors:

Jianqing Fan
Princeton University
371 PUBLICATIONS **48,050** CITATIONS
SEE PROFILE

Gang Li
The Hong Kong Polytechnic University
286 PUBLICATIONS **60,731** CITATIONS
SEE PROFILE

Runze Li
Pennsylvania State University
301 PUBLICATIONS **22,160** CITATIONS
SEE PROFILE

An Overview on Variable Selection for Survival Analysis

Jianqing Fan[1], Gang Li[2], and Runze Li[3]

[1] Jianqing Fan, Department of Operation Research and Financial Engineering, Princeton University, NJ 08544 jqfan@princeton.edu
[2] Gang Li, Department of Biostatistics, University of California Los Angeles, CA 90095-1772 vli@ucla.edu
[3] Runze Li, Department of Statistics, The Pennsylvania State University, University Park, PA16802-2111 rli@stat.psu.edu

Summary. Variable selection are fundamental in high-dimensional statistical modeling. Many authors have proposed various variable selection criteria and procedures for linear regression models (Miller, 2002). Variable selection for survival data analysis poses many challenges because of complicated data structure, and therefore receives much attention in the recent literature. In this article, we will review various existing variable selection procedures for survival analysis. We further propose a unified framework for variable selection in survival analysis via a nonconcave penalized likelihood approach. The nonconcave penalized likelihood approach distinguishes from the traditional variable selection procedures in that it deletes the non-significant covariates by estimating their coefficients as zero. With a proper choice of the penalty function and the regularization parameter, we demonstrate the resulting estimate possesses an oracle property, namely, it performs as well as if the true submodel were known in advance. We further illustrate the methodology by a real data example.

Key words: Accelerated life model, Cox's model, Cox's frailty model, marginal hazards model, variable selection.

2000 Mathematics Subject Classification: 62N01

1 Introduction

Variable selection is vital to survival analysis. In practice, many covariates are often available as potential risk factors. At the initial stage of modeling, data

Contemporary Multivariate Analysis and Experimental Designs—In Celebration of Professor Kaitai Fangs 65th birthday. Edited by Jianqing Fan and Gang Li. The World Scientific Publisher, 2005.

analysts usually introduce a large number of predictors. To enhance model predictability and interpretation, a parsimonous model is always desirable. Thus, selecting significant variables plays crucial roles in model building and is very challenging in the presence of a large number of predictors. Let us first review recent developments of model selection and variable selection for survival data analysis.

Bayesian model selection procedures have been proposed for survival analysis. Faraggi and Simon (1997) and Faraggi (1998) extended the ideas of Lindley (1968) to Cox's proportional hazard models with right censored survival data. To avoid specifying a prior on the baseline hazard function, they use the partial likelihood as the basis for their proposed Bayesian variable selection procedures rather than the full likelihood. Thus, their method indeed is not a proper Bayesian method. Ibrahim, Chen and MacEachern(1999) proposed a full Bayesian variable selection procedure for the Cox model by specifying a nonparametric prior for the baseline function and a parametric prior for the regression coefficients. To implement their methodology, Markov chain Monte Carlo (MCMC) was proposed to compute the posterior model probabilities. Ibrahim and his co-authors (Ibrahim and Chen, 2000, Ibrahim, Chen and Sinha, 2001, Sinha, Chen and Ghosh, 1999) further proposed several Bayesian model assessment criteria. Giudici, Mezzetti and Muliere (2003) proposed a Bayesian nonparametric approach to selecting significant variables in survival analysis based on mixtures of products of Dirichlet process priors. Bayesian variable selection procedures are simple in concept, but hard to implement in high-dimensional modeling due to computational demand for calculating posterior model probabilities.

Most variable selection criteria are closely related to penalized least squares and penalized likelihood. Some traditional variable selection criteria, such as Akaike information criterion (AIC, Akaike, 1974) and Bayesian information criterion (BIC, Schwarz, 1978) can be easily extended to survival analysis. Volinsky and Raftery (2000) extended the BIC to the Cox model. They propose a modification of the penalty term in the BIC so that it is defined in terms of the number of uncensored events instead of the number of observations. Traditional variable selection procedures require subset selection, such as stepwise deletion and the best subset selection. While they are practically useful, subset selection procedures ignore stochastic errors inherited at the stage of variable selections. Hence, their theoretic properties are somewhat difficult to understand. Furthermore, the best subset selection suffers from several drawbacks, the most severe of which is its lack of stability (Breiman, 1996). To retain virtues of the subset selection and to avoid the unstability of the subset selection, Tibshirani (1996) proposed the LASSO variable selection procedures for linear regression models and generalized linear models. The LASSO procedure was further extended to the Cox model in Tibshirani (1997). In an attempt to automatically and simultaneously select variables, Fan and Li (2001) proposed nonconcave penalized approaches for linear regression, robust linear models and generalized linear models, and suggested

the use of smoothly absolute clipped deviation (SCAD) penalty. For simplicity of presentation, we will refer the procedures related to the SCAD penalized likelihood as SCAD. The SCAD is a useful amelioration of LASSO. Fan and Li (2001) demonstrated the SCAD possesses an oracle property, namely, the resulting estimate can correctly identify the true model as if it were known in advance, while the LASSO does not possess this oracle property. Fan and Li (2002) derived a nonconcave penalized partial likelihood for the Cox model and the Cox frailty model, and further illustrate the oracle property of their proposed procedures. In this paper, we aim to provide a unified framework of variable selection for various survival models, including parametric models and the Cox model for univariate survival data, and the Cox frailty model and the marginal hazard model for multivariate failure time.

The paper is organized as follows. In Section 2, we briefly introduce penalized likelihood approaches and extend the nonconcave penalized likelihood approach to parametric models in survival analysis. We derive a penalized partial likelihood procedure for the Cox model using Breslow's "least informative" nonparametric modeling for the cumulative baseline hazard function in Section 3. We extend nonconcave penalized likelihood variable selection procedures to multivariate survival data in Section 4. We deal with some practical implementation issues in Section 5. A real data example in Section 6 is used to illustrate the nonconcave penalized likelihood approach.

2 Nonconcave penalized likelihood approach

2.1 Penalized least squares and penalized likelihood

Most variable selection procedures are related to *penalized least squares*. Suppose that we have the $(d+1)$-dimensional random sample $(\mathbf{x}_i, y_i),\ i = 1, \cdots, n$, from a population $(\mathbf{x}, y)$, where $\mathbf{x}$ is a d-dimensional random vector, and y is a continuous random variable. Consider a linear regression model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i,$$

where $\boldsymbol{\beta}$ is unknown regression coefficients, and ε_i is random error with mean zero and variance σ^2. Define a penalized least squares as

$$Q(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + n\sum_{j=1}^{d} p_{\lambda_{jn}}(|\beta_j|), \tag{1}$$

where $p_{\lambda_{jn}}(\cdot)$ is a given nonnegative penalty function, and λ_{jn}s are regularization parameters, which may depend on n and can be chosen by a data-driven criterion, such as cross-validation (CV) and generalized cross-validation (GCV, Craven and Wahba, 1979). Minimizing (1) yields a penalized least squares estimator. It is worth to note that the penalty functions $p_{\lambda_{jn}}(\cdot)$ in

(4) are not necessarily the same for all j. For example, one may wish to keep important predictors in a parametric model and hence not be willing to penalize their corresponding parameters. For simplicity of presentation, we will assume that the penalty functions for all coefficients are the same, denoted by $p_{\lambda_n}(\cdot)$. Extensions to the case with different thresholding functions do not involve any extra difficulties.

Many variable selection criteria can be derived from the above penalized least squares. Take the penalty function to be the L_0 penalty, namely, $p_{\lambda_n}(|\beta|) = \frac{1}{2}\lambda_n^2 I(|\beta| \neq 0)$, where $I(\cdot)$ is the indicator function. Note that $\sum_{j=1}^{d} I(|\beta_j| \neq 0)$ equals the number of nonzero regression coefficients in the model. Hence many popular variable selection criteria can be derived from (1) with the L_0 penalty by choosing different values of λ_n. For instance, the C_p (Mallows, 1973), AIC (Akaike, 1974), and BIC (Schwarz, 1978) correspond to $\lambda_n = \sqrt{2}(\sigma/\sqrt{n})$, $\sqrt{2}(\sigma/\sqrt{n})$ and $\sqrt{\log n}(\sigma/\sqrt{n})$, respectively, although these criteria were motivated from different principles. Since the L_0 penalty is discontinuous, it requires an exhaustive search over all possible subsets of predictors to find the solution. That is, the algorithm must find the best subset of J predictors for each J in $1, ..., d$, and then choose J to optimize (1). This approach is very expensive in computational cost. Furthermore, the best subset selection suffers from other drawbacks, the most severe of which is its lack of stability as analyzed, for instance, by Breiman (1996).

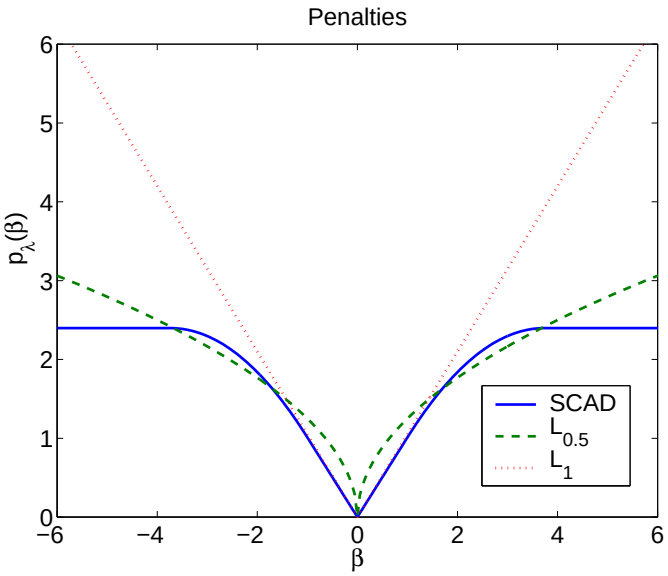

Fig. 1. *Plot of Penalty Functions*

To avoid the drawbacks of the best subset selection, expensive computational cost and the lack of stability, Tibshirani (1996) proposed the LASSO, which can be viewed as the solution of (1) with the L_1 penalty, defined by $p_{\lambda_n}(|\beta|) = \lambda_n|\beta|$. He further demonstrated that LASSO retains the virtues of both best subset selection and ridge regression. Frank and Friedman (1993) considered the L_q penalty, $p_{\lambda_n}(|\beta|) = \lambda_n|\beta|^q$, $(0 < q < 1)$, which yields a "bridge regression". The nonnegative garrote (Breiman, 1995) is in the same spirit as bridge regression. Efron, *et al.*, (2004) further provides deep insights into procedures of the LASSO and the least angle regression. The issue of selection penalty function has been studied in depth by various authors, for instance, Antoniadis and Fan (2001). Fan and Li (2001) suggested the use of the smoothly clipped absolute deviation (SCAD) penalty, defined by

$$p'_{\lambda_n}(\beta) = \lambda_n\{I(\beta \le \lambda_n) + \frac{(a\lambda_n - \beta)_+}{(a-1)\lambda_n} I(\beta > \lambda_n)\} \text{ for some } a > 2 \quad \text{and } \beta > 0,$$

with $p_{\lambda_n}(0) = 0$. This penalty function involves two unknown parameters λ_n and a. Justifying from a Bayesian statistical point of view, Fan and Li (2001) suggested using $a = 3.7$. The Bayes risk cannot be reduced much with other choices of a, and simultaneous data-driven selection of a and λ_n does not have any significant improvements from our experience. Figure 1 depicts the plots of the SCAD, $L_{0.5}$ and L_1 penalty functions.

As shown in Figure 1, the three penalty functions all are singular at the origin. This is a necessary condition for sparsity in variable selection: the resulting estimator automatically sets some small coefficients to be zero (Antoniadis and Fan, 2001). Furthermore, the SCAD and $L_{0.5}$ penalties are nonconvex over $(0, +\infty)$ in order to reduce estimation bias. We refer to penalized least squares with the nonconvex penalties over $(0, \infty)$ as *nonconvex penalized least squares* in order to distinguish from the L_2 penalty, which yields a ridge regression. The SCAD is an improvement over the L_0-penalty in two aspects: saving computational cost and resulting in a continuous solution to avoid unnecessary modeling variation. Furthermore, the SCAD improves bridge regression by reducing modeling variation in model prediction. Although similar in spirit to the L_1-penalty, the SCAD may also improve the L_1-penalty by avoiding excessive estimation bias because the solution of the L_1-penalty could shrink all regression coefficients by a constant, for instance, the soft thresholding rule (Donoho and Johnstone, 1994 and Tibshirani, 1996).

Antoniadis and Fan (2001) and Fan and Li (2001) discussed extensively the choice of the penalty functions. They gave necessary conditions for the penalty function such that penalized least squares estimators to possess the following three desired properties. (i) **Sparsity**: The coefficients of insignificant variables should be estimated as zero. This achieves the purpose of the variable selection. (ii) **Continuity**: The estimated coefficients should be continuous in data to enhance the model stability. This avoids unnecessary variation in the prediction. (iii) **Unbiasedness**: When the true coefficients are large, they should be estimated asymptotically unbiasedly. This avoids unnecessary bi-

ases in the model selection steps. Antoniadis and Fan (2001) and Fan and Li (2001) gave several useful penalty functions that possess these three conditions. This includes the SCAD. Of course, the class of penalty functions satisfied the aforementioned three properties are infinitely many.

The discussion so far has assumed that y is continuous. When the response Y is discrete, such as binary output and count data, generalized linear models (McCullagh and Nelder, 1989) may be used to fit the data. The penalized least squares approach can be adopted to this setting. Conditioning on $\mathbf{x}_i$, suppose that y_i has a density $f_i\{g(\mathbf{x}_i^T\boldsymbol{\beta}), y_i\}$, where g is a known link function. Let $\ell_i = \log f_i$ denote the conditional log-likelihood of y_i. Define a penalized likelihood as

$$\sum_{i=1}^{n} \ell_i(g(\mathbf{x}_i^T\boldsymbol{\beta}), y_i) - n\sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|). \tag{2}$$

Maximizing the penalized likelihood results in a penalized likelihood estimator. The penalized likelihood with a nonconvex penalty over $(0, +\infty)$ is referred to as *nonconcave penalized likelihood*. For certain penalties, such as the SCAD, the selected model based on the nonconcave penalized likelihood satisfies $p_{\lambda_n}(|\beta_j|) = 0$ for certain β_j's. Therefore, model estimation is performed at the same time as model selection. Because the nonconcave penalized likelihood selects variables and estimates parameters simultaneously, this allows us to establish the sampling properties of the resulting estimators. Under certain regularity conditions, Fan and Li (2001) demonstrated how the rates of convergence for the penalized likelihood estimators depend on the regularization parameter λ_n. They further showed that the penalized likelihood estimators perform as well as the oracle procedure in terms of selecting the correct model, when the regularization parameter is appropriately chosen. In practice, a data-driven approach to selecting the regularization parameter is recommended. In Section 5, we present a data-driven method for choosing λ_n using the generalized cross-validation. The optimization of the nonconcave penalized likelihood can be accomplished by the modified Newton-Raphson algorithm with local quadratic approximations (LQA) to the penalty function (Fan and Li 2001). The local quadratic approximation algorithm is also given in Section 5.

2.2 Parametric models in survival data analysis

The penalized likelihood approach can be directly applied for parametric models in survival analysis. Let T, C and $\mathbf{x}$ be respectively the survival time, the censoring time and their associated covariates. Correspondingly, let $Z = \min\{T, C\}$ be the observed time and $\delta = I(T \leq C)$ be the censoring indicator. It is assumed that T and C are conditionally independent given $\mathbf{x}$ and that the censoring mechanism is noninformative. When the observed data $\{(\mathbf{x}_i, Z_i, \delta_i) : i = 1, \cdots, n\}$ is an independently and identically distributed random sample from a certain population $(\mathbf{x}, Z, \delta)$, a complete likelihood of the

data is given by

$$L = \prod_u f(Z_i|\mathbf{x}_i) \prod_c \overline{F}(Z_i|\mathbf{x}_i) = \prod_u h(Z_i|\mathbf{x}_i) \prod_{i=1}^{n} \overline{F}(Z_i|\mathbf{x}_i), \tag{3}$$

where the subscripts c and u denote the product of the censored and uncensored data respectively, and $f(t|\mathbf{x})$, $\overline{F}(t|\mathbf{x})$ and $h(t|\mathbf{x})$ are the conditional density function, the conditional survival function and the conditional hazard function of T given $\mathbf{x}$. Statistical inference in this paper will be based on the likelihood function (3).

In the reminder of this section, we illustrate how to extend the penalized likelihood approach for parametric survival models. Here we focus on accelerated life models, which is one of the most popular parametric life models (Bagdonavičius and Nikulin, 2002). The proposed procedure is ready for applying to other parametric models. The accelerated life models use a linear regression model to fit $\log(T)$, the natural logarithm of T. In other words, the accelerated life models consider

$$\log(T) = \mu + \mathbf{x}^T\boldsymbol{\beta} + \varepsilon. \tag{4}$$

Different choices for the error distribution of ε yields different regression models. Let $\overline{F}_0(t)$ denote the survival function of T when $\mathbf{x} = 0$, i.e., $\overline{F}_0(t)$ is the survival function of $\exp(\mu + \varepsilon)$. Then

$$\overline{F}(t|\mathbf{x}) = P\{T > t|\mathbf{x}\} = \overline{F}_0\{t\exp(-\mathbf{x}^T\boldsymbol{\beta})\}.$$

Furthermore, with $h_0(\cdot)$ being the hazard risk of $\overline{F}_0(\cdot)$,

$$h(t|\mathbf{x}) = h_0\{t\exp(-\mathbf{x}^T\boldsymbol{\beta})\}\exp(-\mathbf{x}^T\boldsymbol{\beta}).$$

Using (3), the log-likelihood of the observed data $\{(\mathbf{x}_i, Z_i, \delta_i) : i = 1, \cdots, n\}$ is

$$\ell_a(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_u \left(-\mathbf{x}_i^T\boldsymbol{\beta} + \log[h_0\{Z_i\exp(-\mathbf{x}_i^T\boldsymbol{\beta})\}]\right) + \sum_{i=1}^{n} \log[\overline{F}_0\{Z_i\exp(-\mathbf{x}_i^T\boldsymbol{\beta})\}], \tag{5}$$

where $\boldsymbol{\theta}$ consists of the unknown parameter involved in the distribution of $\mu + \varepsilon$. Thus, a penalized likelihood for the accelerated life model is

$$\ell_a(\boldsymbol{\beta}, \boldsymbol{\theta}) - n\sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|). \tag{6}$$

Maximizing (6) yields a penalized likelihood estimator for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. With a proper choice of p_{λ_n}, many of estimated coefficients will be zero and hence their corresponding variables do not appear in the model. This achieves the objectives of variable selection.

Example 1. Let the error distribution in (4) be $N(0, \sigma^2)$. This yields a log-normal regression model. Then the survival function of T when $\mathbf{x} = 0$ is

$$\overline{F}_0(t) = 1 - \Phi\left(\frac{\log(t) - \mu}{\sigma}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution of the standard normal distribution. Furthermore, the hazard function when $\mathbf{x} = 0$ is

$$h_0(t) = \frac{\exp\{-(\log(t) - \mu)^2/(2\sigma^2)\}}{t\sqrt{2\pi}\sigma\left\{1 - \Phi\left(\frac{\log(t)-\mu}{\sigma}\right)\right\}}.$$

Plugging $\overline{F}_0(t)$ and $h_0(t)$ into (5), we can derive a closed form for the log-likelihood function. In this example, $\boldsymbol{\theta} = (\mu, \sigma^2)^T$.

Example 2. In this example, we consider the error distribution in (4) to be an extreme value distribution with the following density function

$$f(\varepsilon) = \alpha \exp\{\alpha\varepsilon - \exp(\alpha\varepsilon)\}.$$

The regression model (4) becomes a Weibull regression. By some straightforward calculation, we have

$$\overline{F}_0(t) = \exp(-\nu t^\alpha), \quad \text{and} \quad h_0(t) = \alpha\nu t^{\alpha-1},$$

where $\nu = \exp(-\alpha\mu)$. In this example,

$$h(t|\mathbf{x}) = \alpha\nu t^{\alpha-1}\exp(-\alpha\mathbf{x}^T\boldsymbol{\beta})$$

which is a proportional hazard model. Substituting $\overline{F}_0(t)$ and $h_0(t)$ into (5), the log-likelihood function of the collected data is

$$\ell_a(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_u \{-\alpha\mathbf{x}_i^T\boldsymbol{\beta} + \log(\alpha\nu) + (\alpha - 1)\log(Z_i)\} - \nu\sum_{i=1}^n Z_i^\alpha \exp(-\alpha\mathbf{x}_i^T\boldsymbol{\beta}).$$

where $\boldsymbol{\theta} = (\nu, \alpha)^T$. Maximizing

$$\ell_a(\boldsymbol{\beta}, \boldsymbol{\theta}) - n\sum_{j=1}^d p_{\lambda_n}(|\beta_j|).$$

yields a penalized maximum likelihood estimate for $(\boldsymbol{\beta}, \boldsymbol{\theta})$.

Example 3. Take the error distribution in (4) to be logistic distribution with density

$$f(\varepsilon) = \frac{\exp(\alpha\varepsilon)}{\sigma\{1 + \exp(\alpha\varepsilon)\}^2}.$$

Then model (4) becomes a log-logistic regression model. It follows that

$$\overline{F}_0(t) = \frac{1}{1+\nu t^\alpha}, \quad \text{and} \quad h_0(t) = \frac{\alpha\nu t^{\alpha-1}}{1+\nu t^\alpha},$$

where $\nu = \exp(-\alpha\mu)$. In this example,

$$h(t|\mathbf{x}) = \frac{\alpha\nu t^{\alpha-1}\exp(-\alpha\mathbf{x}^T\boldsymbol{\beta})}{1+\nu t^\alpha \exp(-\alpha\mathbf{x}^T\boldsymbol{\beta})}.$$

A closed form for the log-likelihood function can be derived by using the explicit expression of $\overline{F}_0(t)$ and $h_0(t)$. In this example, $\boldsymbol{\theta} = (\lambda, \alpha)^T$.

3 Variable selection for Cox's models

The Cox proportional hazard model assumes

$$h(t|\mathbf{x}) = h_0(t)\exp(\mathbf{x}^T\boldsymbol{\beta}), \tag{7}$$

where the baseline hazard function $h_0(t)$ is an unspecified function. To present explicitly the likelihood function of the observed data $\{(\mathbf{x}_i, Z_i, \delta_i) : i = 1, \cdots, n\}$ from Cox's proportional hazards model, more notation is needed. Let $t_1^0 < \cdots < t_N^0$ denote the ordered observed failure times. Let (j) provide the label for the item falling at t_j^0 so that the covariates associated with the N failures are $\mathbf{x}_{(1)}, \cdots, \mathbf{x}_{(N)}$. Let R_j denote the risk set right before the time $t_j^0 : R_j = \{i : Z_i \geq t_j^0\}$. The likelihood in (3) becomes

$$L = \prod_{i=1}^{N} h_0(Z_{(i)})\exp(\mathbf{x}_{(i)}^T\boldsymbol{\beta}) \prod_{i=1}^{n} \exp\{-H_0(Z_i)\exp(\mathbf{x}_i^T\boldsymbol{\beta})\},$$

where $H_0(\cdot)$ is the cumulative baseline hazard function. The corresponding penalized log-likelihood function is

$$\sum_{i=1}^{N}[\log\{h_0(Z_{(i)})\} + \mathbf{x}_{(i)}^T\boldsymbol{\beta}] - \sum_{i=1}^{n}\{H_0(Z_i)\exp(\mathbf{x}_i^T\boldsymbol{\beta})\} - n\sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|). \tag{8}$$

Since the baseline hazard and cumulative hazard functions are unknown and have not been parameterized, the penalized log-likelihood function (8) is not ready for optimization yet. Following Breslow's idea, consider the "least informative" nonparametric modeling for $H_0(\cdot)$, in which $H_0(t)$ has a possible jump h_j at the observed failure time t_j^0. More precisely, let $H_0(t) = \sum_{j=1}^{N} h_j I(t_j^0 \leq t)$. Then

$$H_0(Z_i) = \sum_{j=1}^{N} h_j I(i \in R_j). \tag{9}$$

Using (9), the logarithm of penalized likelihood function of (8) becomes

$$\sum_{j=1}^{N}\{\log(h_j)+\mathbf{x}_{(j)}^T\boldsymbol{\beta}\}-\sum_{i=1}^{n}\{\sum_{j=1}^{N}h_jI(i\in R_j)\exp(\mathbf{x}_i^T\boldsymbol{\beta})\}-n\sum_{j=1}^{d}p_{\lambda_n}(|\beta_j|). \quad (10)$$

Taking the derivative with respect to h_j and setting it to be zero, we obtain that

$$\hat{h}_j=\{\sum_{i\in R_j}\exp(\mathbf{x}_i^T\boldsymbol{\beta})\}^{-1}. \quad (11)$$

Substituting $\hat{h}_j$ into (10), we get the penalized partial likelihood

$$\sum_{j=1}^{N}[\mathbf{x}_{(j)}^T\boldsymbol{\beta}-\log\{\sum_{i\in R_j}\exp(\mathbf{x}_i^T\boldsymbol{\beta})\}]-n\sum_{j=1}^{d}p_{\lambda_n}(|\beta_j|)\stackrel{\text{def}}{=}\ell_c(\boldsymbol{\beta})-n\sum_{j=1}^{d}p_{\lambda_n}(|\beta_j|), \quad (12)$$

after dropping a constant term "$-N$". When $p_\lambda(\cdot)\equiv 0$, (12) is the partial likelihood function (Cox, 1975). Thus, the penalized likelihood indeed is the penalized partial likelihood. The penalized likelihood estimate of $\boldsymbol{\beta}$ is obtained via maximizing (12) with respect to $\boldsymbol{\beta}$.

Numerical comparison in Fan and Li (2002) shows that the SCAD performs as well as the oracle estimate, and outperforms the penalized likelihood with the L_1 penalty and the best subset selection with the BIC. This oracle property is further demonstrated by the following asymptotic formulation.

Let $\boldsymbol{\beta}_0$ be the true value of $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_0=(\beta_{10},\cdots,\beta_{d0})^T=(\boldsymbol{\beta}_{10}^T,\boldsymbol{\beta}_{20}^T)^T$. Without loss of generality, assume that $\boldsymbol{\beta}_{20}=\mathbf{0}$. Denote by s the number of the component of $\boldsymbol{\beta}_1$. Fan and Li (2002) first showed that under certain regularity conditions, if $\lambda_n\to 0$, then there exists a local maximizer $\hat{\boldsymbol{\beta}}$ of the SCAD penalized partial likelihood function in (12) such that $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|=O_P(n^{-1/2})$. They further proved that if $\lambda_n\to 0$ and $\sqrt{n}\lambda_n\to\infty$, then under certain regularity conditions, with probability tending to 1, the root n consistent local maximizer $\hat{\boldsymbol{\beta}}=(\hat{\boldsymbol{\beta}}_1^T,\hat{\boldsymbol{\beta}}_2^T)^T$ of the penalized partial likelihood in (12) with the SCAD penalty must satisfy

(i) (**Sparsity**) $\hat{\boldsymbol{\beta}}_2=\mathbf{0}$;
(ii) (**Asymptotic normality**)

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10})\to N(\mathbf{0},I_1^{-1}(\boldsymbol{\beta}_{10})),$$

where $I_1(\boldsymbol{\beta}_{10})$ is the first $s\times s$ submatrix of $I(\boldsymbol{\beta}_0)$, the Fisher information matrix of the partial likelihood.

Property (i) and (ii) is referred to as an oracle property, which provides a foundation for variable selection. The sparsity (i) indicates that $\hat{\boldsymbol{\beta}}_2=0$ is the same as the oracle estimator who knows in advance $\boldsymbol{\beta}_2=0$. Furthermore, the estimator $\hat{\boldsymbol{\beta}}_1$ shares the same sampling property as the oracle estimator,

and is more efficient than the maximum partial likelihood estimator (without penalty). In other words, the SCAD possesses this oracle property. The oracle property holds not only for the SCAD, but also for a class of infinitely many penalty functions. But it does not hold for the L_1 penalty due to the excessive biases inherent to the L_1 penalty. As demonstrated in Cai, Fan, Li and Zhou (2003), the oracle property is also valid for the setting in which the number of covariates is allowed to depend on n and the number of nonzero coefficients, say s_n, tends to infinite as $n \to \infty$. See Fan and Peng (2004) for a formulation under general settings.

4 Variable selection for multivariate survival data

It is assumed for the Cox proportional hazards model that the survival times of subjects are independent. This assumption might be violated in some situations, in which the collected data are correlated. The well-known Cox model (Cox, 1972) is not valid in this situation because independence assumption among individuals is violated. Extensions of the Cox regression model to the analysis of multivariate failure time data include frailty model and marginal model. In this section, we extend the nonconcave penalized likelihood approach for the frailty model and the marginal model.

4.1 Frailty models

One popular approach to modeling correlated survival times is to use a frailty model. A frailty corresponds to a random block effect that acts multiplicatively on the hazard rates of all subjects in a group. In this section, we only consider the Cox proportional hazard frailty model, in which it is assumed that the hazard rate for the j-th subject in the i-th subgroup is

$$h_{ij}(t|\mathbf{x}_{ij}, u_i) = h_0(t)u_i \exp(\mathbf{x}_{ij}^T\boldsymbol{\beta}),\ i = 1, \cdots, n, j = 1, \cdots, J_i, \qquad (13)$$

where the u_i's are associated with frailties, and they are a random sample from some population. It is frequently assumed that given the frailty u_i, the data in the i-th group are independent. The most frequently used distribution for frailty is the gamma distribution due to its simplicity. Assume without loss of generality that the mean of frailty is 1 so that all parameters involved are estimable. For the gamma frailty model, the density of u is

$$g(u) = \frac{\alpha^\alpha u^{\alpha-1}\exp(-\alpha\, u)}{\Gamma(\alpha)}.$$

From (3), the full likelihood of "pseudo-data" $\{(u_i, \mathbf{x}_{ij}, Z_{ij}, \delta_{ij}) : i = 1, \cdots, n, j = 1 \cdots, J_i\}$ is

$$\prod_{i=1}^{n}\prod_{j=1}^{J_i}[\{h(z_{ij}|\mathbf{x}_{ij}, u_i)\}^{\delta_{ij}}\overline{F}(z_{ij}|\mathbf{x}_{ij}, u_i)]\prod_{i=1}^{n} g(u_i).$$

Integrating the full likelihood function with respect to $u_1, \cdots, u_n$, the likelihood of the observed data is given by

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}) = \exp\{\boldsymbol{\beta}^T(\sum_{i=1}^{n}\sum_{j=1}^{J_i}\delta_{ij}\mathbf{x}_{ij})\}\prod_{i=1}^{n}\frac{\alpha^\alpha\prod_{j=1}^{J_i}\{h_0(z_{ij})\}^{\delta_{ij}}}{\Gamma(\alpha)\{\sum_{j=1}^{J_i}H_0(z_{ij})\exp(\mathbf{x}_{ij}^T\boldsymbol{\beta})+\alpha\}^{(A_i+\alpha)}}, \tag{14}$$

where $\boldsymbol{\theta} = (\alpha, H)$, and $A_i = \sum_{j=1}^{J_i}\delta_{ij}$. The log-likelihood of the observed data is

$$\begin{aligned}\ell_f(\boldsymbol{\beta}, \boldsymbol{\theta}) = &\sum_{i=1}^{n}\{\sum_{j=1}^{J_i}\delta_{ij}\log h(z_{ij}) - [(A_i+\alpha)\log\{\sum_{j=1}^{J_i}H_0(z_{ij})\exp(\mathbf{x}_{ij}^T\boldsymbol{\beta})+\alpha\}]\}\\ &+\sum_{i=1}^{n}\{\boldsymbol{\beta}^T(\sum_{j=1}^{J_i}\delta_{ij}\mathbf{x}_{ij}) + \alpha\log\alpha - \log\Gamma(\alpha)\}\end{aligned}$$

Therefore the logarithm of the penalized likelihood of the observed data is

$$\ell_f\{\boldsymbol{\beta}, h(\cdot)\} - n\sum_{j=1}^{d}p_\lambda(|\beta_j|). \tag{15}$$

To eliminate the nuisance parameter $h(\cdot)$, we again employ the profile likelihood method. Consider the "least informative" nonparametric modeling for $H_0(\cdot)$:

$$H_0(z) = \sum_{l=1}^{N}\lambda_l I(z_l \le z), \tag{16}$$

where $\{z_1, \cdots, z_N\}$ are pooled observed failure times.

Substituting (16) into (15), then differentiating it with respect to λ_l, $l = 1, \cdots, N$, the root of the corresponding score function should satisfy the following equations:

$$\lambda_l^{-1} = \sum_{i=1}^{n}\frac{(A_i+\alpha)\sum_{j=1}^{J_i}I(z_l \le z_{ij})\exp(\mathbf{x}_{ij}^T\boldsymbol{\beta})}{\sum_{k=1}^{N}\lambda_k\sum_{j=1}^{J_i}I(z_k \le z_{ij})\exp(\mathbf{x}_{ij}^T\boldsymbol{\beta})+\alpha} \qquad \text{for } l = 1, \cdots, N. \tag{17}$$

The above solution does not admit a close form, neither does the profile likelihood function. However, the maximum profile likelihood can be implemented as follows. With initial values for $\alpha, \boldsymbol{\beta}$ and λ_l, update $\{\lambda_l\}$ from (17) and obtain the penalized profile likelihood of (15). With known $H_0(\cdot)$ defined by (16), maximize the penalized likelihood (15) with respect to $(\alpha, \boldsymbol{\beta})$, and iterate between these two steps. When the Newton-Raphson algorithm is applied to the penalized likelihood (15), it involves the first two order derivatives of the gamma function, which may not exist for certain value of α. One approach to avoid this difficulty is the use of a grid of possible values for the frailty parameter α and finding the maxima over this discrete grid, as

suggested by Nielsen *et al.* (1992). Our simulation experience shows that the estimate of $\boldsymbol{\beta}$ is quite empirically robust to the chosen grid of possible values for α. This profile likelihood method even without the task of variable selection provides a viable alternative approach to the EM algorithm frequently used in the frailty model.

A natural initial estimator for $\boldsymbol{\beta}$ is the maximum pseudo-partial likelihood estimates of $\boldsymbol{\beta}$ ignoring possible dependency within each group. The corresponding $h_1, \cdots, h_N$ in (11) may serve as an initial estimator for $\lambda_1, \cdots, \lambda_N$. Hence given a value of α and initial values of $\boldsymbol{\beta}$ and $\lambda_1, \cdots, \lambda_N$, update the values of $\lambda_1, \cdots, \lambda_N$ and α, $\boldsymbol{\beta}$ in turn until they converge or the penalized profile likelihood fails to change substantially. The proposed algorithm avoids optimizing a high-dimensional problem. It will give us an efficient estimate for $\boldsymbol{\beta}$. The algorithm may converge slowly or even not converge. In this situation, the idea of one-step estimator (see Bickel, 1975) provides us an alternative approach.

Fan and Li (2002) assessed the finite sample performance of the resulting estimate by extensive Monte Carlo simulation. From their numerical comparisons, it can be seen that the SCAD performs almost as well as the oracle estimator in terms of model error, and it outperforms the penalized likelihood with the L_1 penalty in terms of model complexity and model error. The performance of the SCAD is similar to the best subset selection with BIC in terms of model complexity and model error, but the computational time of SCAD is dramatically less than that of the best subset selection. Under certain regularity conditions, Fan and Li (2002) showed that if $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$, then the resulting estimate of the SCAD is root n consistent, and with probability tending to one, $\hat{\boldsymbol{\beta}}_2 = 0$ and

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}) \to N\left\{\mathbf{0}, \tilde{I}_1^{-1}(\boldsymbol{\theta}_{10})\right\},$$

where $\tilde{I}_1(\boldsymbol{\theta}_{10})$ consists of the first $(s+1) \times (s+1)$ submatrix of $\tilde{I}_0(\boldsymbol{\theta}_{10}, \mathbf{0})$, the Fisher information matrix of the frailty model, and $\hat{\boldsymbol{\theta}}_1 = (\hat{\alpha}, \hat{\boldsymbol{\beta}}_1^T)^T$, $\boldsymbol{\theta}_{10} = (\alpha_0, \boldsymbol{\beta}_{10}^T)^T$.

4.2 Marginal Hazard Models

The interpretations of the regression coefficients in the frailty model are different from those in the Cox model. Consequently, when the correlation among the observations is not of interest, the marginal proportional hazard models have received much attention in the recent literature because they are semiparametric models and retain the virtue of the Cox model (e.g., Wei, Lin and Weissfeld 1989, Lee, Wei and Amato 1992, Liang, Self and Chang 1993, Lin 1994, Cai and Prentice 1995, 1997, Cai, 1997, Spiekerman and Lin 1998 and Clegg, Cai and Sen 1999 among others).

To fix notation, suppose that there are n independent clusters and each cluster has K subjects. For each subject, J types of failure may occur. For

the failure time in the case of the jth type of failure on subject k in cluster i, the marginal mixed baseline hazards model is taken as

$$h_{ijk}(t|\mathbf{x}_{ijk}(t)) = h_{0j}(t)\exp\{\boldsymbol{\beta}^T\mathbf{x}_{ijk}(t)\}, \tag{18}$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_d)^T$ is a vector of unknown regression coefficients, $\mathbf{x}_{ijk}(t)$ is a possibly external time-dependent covariate vector, and $h_{0j}(t)$ and $h_0(t)$ are unspecified baseline hazard functions.

The marginal model approach does not specify correlation structure for the failure times within a cluster, hence inferences are based on a pseudo-partial likelihood approach. Under a working independence assumption (Wei, Lin and Weissfield 1989), i.e., assuming the independence among failure times in a cluster, we obtain the logarithm of a pseudo-partial likelihood function of the observed data $\{(\mathbf{x}_{ijk}, Z_{ijk}, \delta_{ijk}) : i = 1, \cdots, n, j = 1 \cdots, J, k = 1, \cdots, K\}$ from model (18) as following:

$$\begin{aligned}\ell_p(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{J}\sum_{k=1}^{K}\delta_{ijk}\Big(&\boldsymbol{\beta}^T\mathbf{x}_{ijk}(Z_{ijk}) \\ &- \log\left[\sum_{l=1}^{n}\sum_{g=1}^{K}Y_{ljg}(Z_{ijk})\exp\{\boldsymbol{\beta}^T\mathbf{x}_{ljg}(Z_{ijk})\}\right]\Big).\end{aligned} \tag{19}$$

where $Y(t) = I(Z \geq t)$ be the at-risk indicator. We use a penalized pseudo-partial likelihood for model (18) which is defined as

$$\mathcal{L}(\boldsymbol{\beta}) = \ell_p(\boldsymbol{\beta}) - n\sum_{j=1}^{d}p_{\lambda_n}(|\beta_j|). \tag{20}$$

Let

$$a_n = \max\{|p'_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}, \quad \text{and} \quad b_n = \max\{|p''_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}. \tag{21}$$

We first show that there exists a penalized pseudo-partial likelihood estimator that converges at rate $O_P(n^{-1/2} + a_n)$, and then establish the oracle property for the resulting estimator. We only state the main theoretic results here and leave the regularity conditions in the Appendix. Technical proofs are given in Cai, Fan, Li and Zhou (2003).

Theorem 1. *Under Conditions A-D in the Appendix, if both a_n and b_n tend to 0 as $n \to \infty$, then with probability tending to one, there exists a local maximizer $\hat{\boldsymbol{\beta}}$ of $\mathcal{L}(\boldsymbol{\beta})$ defined in (6) such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2} + a_n)$.*

From Theorem 1, provided that $a_n = O(n^{-1/2})$, which can be achieved by choosing proper $\lambda'_n s$, there exists a root n consistent penalized pseudo-partial likelihood estimator. Denote by

$$\Sigma = \text{diag}\{p''_{\lambda_n}(|\beta_{10}|), \cdots, p''_{\lambda_n}(|\beta_{s0}|)\},$$

and

$$\mathbf{b} = (p'_{\lambda_n}(|\beta_{10}|)\text{sgn}(\beta_{10}), \cdots, p'_{\lambda_n}(|\beta_{s0}|)\text{sgn}(\beta_{s0})).$$

Theorem 2. *Assume that the penalty function $p_{\lambda_n}(|\beta_j|)$ satisfies that*

$$\liminf_{n\to\infty} \liminf_{\beta\to 0+} p'_{\lambda_n}(\beta)/\lambda_n > 0. \tag{22}$$

If $\lambda_n \to 0$, $\sqrt{n}\lambda_n \to \infty$ and $a_n = O(n^{-1/2})$, then under the conditions of Theorem 1, with probability tending to 1, the root n consistent local maximizer $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ in Theorem 1 must satisfy that $\hat{\boldsymbol{\beta}}_2 = 0$, and

$$\sqrt{n}\{A_{11} + \Sigma\}\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (A_{11} + \Sigma)^{-1}\mathbf{b}\} \to N(0, D_{11}) \tag{23}$$

in distribution, where A_{11} and D_{11} consist of the first s columns and rows of $A(\boldsymbol{\beta}_{10}, \mathbf{0})$ and $D(\boldsymbol{\beta}_{10}, \mathbf{0})$ defined in the Appendix, respectively.

5 Practical implementation

5.1 Local quadratic approximation and standard error formula

The L_q, $(0 < q \leq 1)$, and SCAD penalty functions are singular at the origin, and they do not have continuous second order derivatives. Therefore, maximizing the nonconcave penalized likelihood is challenging. Fan and Li (2001) proposed a unified algorithm for their nonconcave penalized likelihood using a local quadratic approximation for the penalty function. The unified algorithm is ready for the penalized likelihood function

$$Q(\boldsymbol{\beta}, \boldsymbol{\theta}) = \ell(\boldsymbol{\beta}, \boldsymbol{\theta}) - n\sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|). \tag{24}$$

where $\ell(\boldsymbol{\beta}, \boldsymbol{\theta})$ may be the likelihood function $\ell_a(\boldsymbol{\beta}, \boldsymbol{\theta})$ in Section, $\ell_c(\boldsymbol{\beta})$ for the Cox model, $\ell_f(\boldsymbol{\beta}, \boldsymbol{\theta})$ for the Cox's frailty model and $\ell_p(\boldsymbol{\beta})$ for the marginal model. Although the penalty function is singular at the origin and may not have continuous 2nd order derivative. Fan and Li (2001) propose to locally approximate using a quadratic function as follows. Set the initial value to be the maximum likelihood estimate (without penalty). Under certain regularity conditions, the MLE is root n consistent, and therefore it is close to the true value. Suppose that we are given an initial value $\boldsymbol{\beta}^0$ that is close to the minimizer of (24). If β_j^0 is very close to 0, then set $\hat{\beta}_j = 0$. Otherwise they can be locally approximated by a quadratic function as

$$[p_{\lambda_n}(|\beta_j|)]' = p'_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j) \approx \{p'_{\lambda_n}(|\beta_j^0|)/|\beta_j^0|\}\beta_j,$$

when $\beta_j \neq 0$. In other words,

$$p_{\lambda_n}(|\beta_j|) \approx p_{\lambda_n}(|\beta_j^0|) + \frac{1}{2}\{p'_{\lambda_n}(|\beta_j^0|)/|\beta_j^0|\}(\beta_j^2 - \beta_j^{02}), \text{ for } \beta_j \approx \beta_j^0. \quad (25)$$

With the aid of the quadratic approximation, the maximization of $Q(\boldsymbol{\beta}, \boldsymbol{\theta})$ can be carried out by using the Newton-Raphson algorithm.When the algorithm converges, the estimator satisfies the condition

$$\frac{\partial \ell(\hat{\boldsymbol{\beta}}^0)}{\partial \beta_j} + np'_{\lambda_n}(|\hat{\beta}_j^0|)\text{sgn}(\hat{\beta}_j^0) = 0,$$

the penalized likelihood equation, for non-zero elements of $\hat{\boldsymbol{\beta}}^0$.

Following conventional techniques in the likelihood setting, we can estimate the standard error of the resulting estimate by using the sandwich formula. Specifically, the corresponding sandwich formula can be used as an estimator for the covariance of the estimates $\hat{\boldsymbol{\beta}}_1$, the non-vanishing component of $\hat{\boldsymbol{\beta}}$. That is,

$$\{\nabla^2 \ell(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\theta}}) + n\Sigma_\lambda(\hat{\boldsymbol{\beta}}_1)\}^{-1}\widehat{\text{cov}}\{\nabla \ell(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\theta}})\}\{\nabla^2 \ell(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\theta}}) - n\Sigma_\lambda(\hat{\boldsymbol{\beta}}_1)\}^{-1}, \quad (26)$$

where

$$\Sigma_\lambda(\hat{\boldsymbol{\beta}}_1) = \text{diag}\{p'_\lambda(|\hat{\beta}_1|)/|\hat{\beta}_1|, \cdots, p'_\lambda(|\hat{\beta}_s|)/|\hat{\beta}_s|, 0, \cdots, 0\},$$

where the number of zeros equals the dimension of $\boldsymbol{\theta}$, and s the dimension of $\hat{\boldsymbol{\beta}}_1$.

5.2 Selection of regularization parameters

To implement the methods described in previous sections, it is desirable to have an automatic method for selecting the thresholding parameter λ involved in $p_\lambda(\cdot)$ based on data. Here we estimate λ via minimizing an approximate generalized cross-validation (GCV) statistic (Craven and Wahba, 1979). By some straightforward calculation, the effective number of parameters for $Q(\boldsymbol{\beta}, \boldsymbol{\theta})$ in (24) in the last step of the Newton-Raphson algorithm iteration is

$$e(\lambda) = \text{tr}[\{\nabla^2 \ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) + \Sigma_\lambda(\hat{\boldsymbol{\beta}})\}^{-1}\nabla^2 \ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})].$$

Therefore the generalized cross-validation statistic is defined by

$$\text{GCV}(\lambda) = \frac{-\ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})}{n\{1 - e(\lambda)/n\}^2}$$

and $\hat{\lambda} = \text{argmin}_\lambda\{\text{GCV}(\lambda)\}$ is selected. The minimization can be carried out by searching over a grid of points for λ.

6 An Example

We illustrate the proposed variable selection procedures by an analysis of a data set collected in the Framingham Heart Study (FHS, Dawber, 1980). In this study, multiple failure outcomes, for instance, times to coronary heart disease (CHD) and cerebrovascular accident (CVA), were observed from the same individual. In addition, as the primary sampling unit was the family, failure times recorded are likely to be dependent for the individuals within a family.

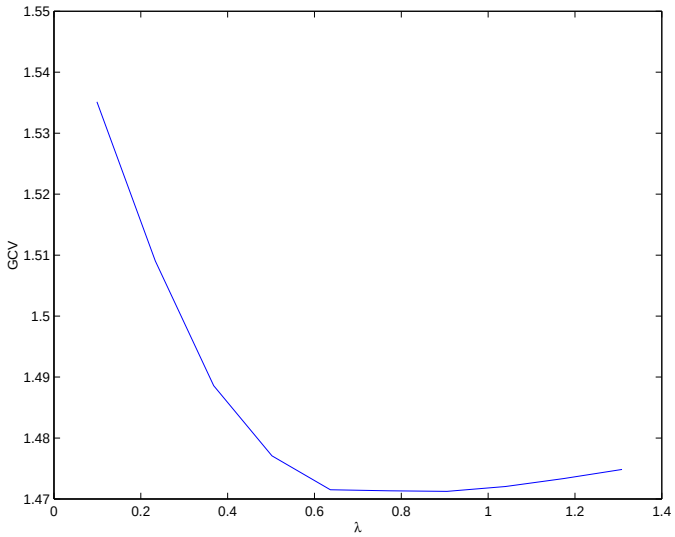

Fig. 2. *Plot of Generalized Cross-Validation for the Framingham Heart Study Analysis*

For simplicity, we consider only time to obtain first evidence of CHD and of CVA, and analyze only data for participants in the FHS study who had an examination at age 44 or 45 and were disease-free at that examination. By disease-free we mean that there exists no history of hypertension or glucose intolerance and no previous experience of a CHD or CVA. The time origin is the time of the examination at which an individual participated in the study and the follow up information is up to year 1980. The risk factors of interest are: body mass index (BMI), denoted by x_1, cholesterol level (x_2), systolic blood pressure (x_3), smoking status (x_4), coded by 1 if this individual is smoking, and 0 otherwise, gender (x_5), coded by 1 for female and 0 for male. The values of risk factors were taken from the biennial examination at which an individual was included in the sample. Because some individuals were in

the study several years prior to inclusion into the data set, the waiting time, denoted by x_6, from entering the study to reaching 44 or 45 years of age was used as a covariate to account for the cohort effect. Since x_1, x_2, x_3 and x_6 are continuous covariates, they are standardized individually in our analysis.

Table 1. Estimated Coefficients and Standard Errors for the FHS data

	CHD	CVA
Effect	$\hat{\beta}$(SE($\hat{\beta}$))	$\hat{\beta}$(SE($\hat{\beta}$))
x_1	0.0810 (0.1288)	0.4773 (0.2423)
x_2	0.0576 (0.1200)	-0.2409 (0.2655)
x_3	0.4129 (0.1570)	0.2917 (0.1477)
x_4	0.4754 (0.2361)	0.7077 (0.3587)
x_5	-0.3666 (0.2543)	-0.1016 (0.2890)
x_6	0.0249 (0.0802)	-0.1395 (0.1916)
x_1^2	-0.0743 (0.0512)	0 (−)
x_2^2	0 (−)	-0.0768 (0.1052)
x_3^2	0 (−)	0 (−)
x_6^2	0 (−)	0.2062 (0.1229)
$x_1 * x_2$	0 (−)	0 (−)
$x_1 * x_3$	0 (−)	-0.2224 (0.1435)
$x_1 * x_4$	0.1409 (0.1495)	-0.2207 (0.2628)
$x_1 * x_5$	0 (−)	0 (−)
$x_1 * x_6$	-0.1060 (0.0808)	0 (−)
$x_2 * x_3$	0 (−)	0 (−)
$x_2 * x_4$	0.1550 (0.1425)	0.5702 (0.3766)
$x_2 * x_5$	0 (−)	0 (−)
$x_2 * x_6$	0 (−)	0 (−)
$x_3 * x_4$	-0.1952 (0.1489)	0 (−)
$x_3 * x_5$	-0.2054 (0.1378)	0 (−)
$x_3 * x_6$	0 (−)	0 (−)
$x_4 * x_5$	-0.3071 (0.3106)	0 (−)
$x_4 * x_6$	0 (−)	0 (−)
$x_5 * x_6$	0 (−)	0.5753 (0.2545)

To explore possible nonlinear effects and interaction effects of the risk factors, we include all main effects, quadratic effects and interaction effects of the risk factors and covariates in the full model. This results in a mixed baseline hazard model with 50 covariates:

$$h_{ijk}(t, \mathbf{x}_{ijk}) = h_{0j}(t) \exp\{\boldsymbol{\beta}_j^T \mathbf{x}_{ijk}\}, \tag{27}$$

where $\mathbf{x}_{ijk}$ consists of all possible linear, quadratic and interaction terms of the risk factors and covariates x_1 to x_6. Model (27) allows different baseline hazards and different regression coefficients for CHD and CVA, but an identical baseline hazards for siblings. A thorough analysis of this data set was also given in Cai, Fan, Li and Zhou (2003).

The maximum pseudo-partial likelihood estimate for $\boldsymbol{\beta}$ is computed. The natural logarithm of the pseudo-partial likelihood for the full model of 50 covariates is -2017.9590. Next we apply the SCAD procedure to model (7) to select significant variables. In the implementation of the SCAD procedure, since all covariates are important confounding variables, we included them in the model by not penalizing the linear main effect of x_1 to x_6. Thus, all linear effects are included in the selected model. The GCV method is used to select the regularization parameter. Figure 2 depicts the plot of GCV score versus λ. The regularization parameter λ equals 0.9053, selected by minimizing the GCV scores. The resulting estimate and standard error for $\boldsymbol{\beta}$ in the selected model is depicted in Table 1. The logarithm of the pseudo-partial likelihood for the model selected by the SCAD with the selected λ is -2022.6635. This represents an increase of the logarithm of the pseudo-partial likelihood by 10.1923 from that of the full model, which is much less than 25, the number of covariates excluded from the full model. From extension of Theorem 3 of Cai (1999), the limiting distribution of the pseudo-partial likelihood ratio statistic is a weighted sum of Chi-square distributions with 1 degree of freedom. Based on 100,000 Monte Carlo simulations, we computed the p-value, which equals 0.9926. This is in favor of the selected model. We further compared the selected model by SCAD with the one selected by the naive approach. The corresponding pseudo-partial likelihood ratio statistic is 90.4989 with p-value 0.0000 obtained by 100,000 Monte Carlo simulations. This is also in favor of the selected model by SCAD.

In another confirmation of the selected model, we compare the selected model with the linear main effect model which include only all the linear main effects of x_1 to x_6. The maximum pseudo-partial likelihood estimate for the unknown regression coefficients is computed, and the natural logarithm of the pseudo-partial likelihood for the linear main effect model is -2034.6527. The pseudo-partial likelihood ratio statistic for testing H_0: the linear main effect model versus H_1: the selected model, is 23.9783. Based on 100,000 Monte Carlo simulations, the corresponding p-value equals 0.0353. This indicates that the selected model fits the data better than the model with only the linear main effects.

Acknowledgements

Fan's research was supported by an NSF grant DMS-0354223 and an NIH grant R01 HL69720. G. Li's research was partially supported by NIH grant CA78314-03. R. Li's research was supported by an NSF grant DMS-0348869 and a National Institute on Drug Abuse (NIDA) Grant 1-P50-DA10075.

Appendix: Regularity Conditions

To facilitate the notation, let $N_{ijk}(t) = I(Z_{ijk} \le t, \delta_{ijk} = 1)$ be the counting process, and $h_{ijk}(t)$ and $M_{ijk}(t) = N_{ijk}(t) - \int_0^t Y_{ijk}(u)h_{ijk}(u)\,du$ be their corresponding marginal hazards function and marginal martingale, respectively, with respect to the filtration $\mathcal{F}_{jk}(t^-)$, where $\mathcal{F}_{jk}(t)$ is the σ-field generated by $\{N_{ijk}(u), Y_{i11}(u), ..., Y_{iJK}(u), \mathbf{x}_{i11}(u), ..., \mathbf{x}_{iJK}(u);\, 0 \le u \le t, i = 1, ..., n\}$. Define

$$\begin{aligned}
\mathbf{S}_{jk}^{(d)}(\boldsymbol{\beta}; t) &= \frac{1}{n}\sum_{i=1}^{n} Y_{ijk}(t)\mathbf{x}_{ijk}(t)^{\otimes d}\exp\{\boldsymbol{\beta}^T\mathbf{x}_{ijk}(t)\}, \quad d = 0, 1, 2\\
\mathbf{S}_{j\cdot}^{(d)}(\boldsymbol{\beta}; t) &= \sum_{k=1}^{K}\mathbf{S}_{jk}^{(d)}(\boldsymbol{\beta}; t), \quad d = 0, 1, 2,\\
\mathbf{E}_j(\boldsymbol{\beta}; t) &= \mathbf{S}_{j\cdot}^{(1)}(\boldsymbol{\beta}; t)/\mathbf{S}_{j\cdot}^{(0)}(\boldsymbol{\beta}; t),\\
\mathbf{V}_j(\boldsymbol{\beta}; t) &= \mathbf{S}_{j\cdot}^{(2)}(\boldsymbol{\beta}, t)/\mathbf{S}_{j\cdot}^{(0)}(\boldsymbol{\beta}, t) - \mathbf{E}_j(\boldsymbol{\beta}; t)^{\otimes 2},
\end{aligned}$$

where $\mathbf{a}^{\otimes 0} = 1$, $\mathbf{a}^{\otimes 1} = 1$, and $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$ for a vector $\mathbf{a}$.

Regularity conditions:

(A) For simplicity, assume that T_{ijk} takes values on a finite interval $[0, \tau]$, and $\int_0^\tau h_{0j}(t)\,dt < \infty$ for $j = 1, \cdots, J$.

(B) There exists a neighborhood $\mathcal{B}$ of the true value $\boldsymbol{\beta}_0$ that satisfies each of the following conditions: (1) there exists a scalar, vector, and matrix function $\mathbf{s}_{jk}^{(d)}(\boldsymbol{\beta}, t)$ $(d = 0, 1, 2)$ defined on $\mathcal{B} \times [0, \tau]$ such that $\sup_{t\in[0,\tau],\boldsymbol{\beta}\in\mathcal{B}} \|\mathbf{S}_{jk}^{(d)}(\boldsymbol{\beta}, t) - \mathbf{s}_{jk}^{(d)}(\boldsymbol{\beta}, t)\| \to 0$ in probability; (2) there exists a matrix $\mathbf{D} = \mathbf{D}(\boldsymbol{\beta})$ such that

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{var}(\mathbf{D}_i) \to \mathbf{D},$$

where

$$\mathbf{D}_i = \sum_{j=1}^{J}\sum_{k=1}^{K}\int_0^\tau \{\mathbf{x}_{ijg}(t) - \mathbf{e}_j(\boldsymbol{\beta}_0; t)\}\,dM_{ijk}(t),$$

and $\mathbf{e}_j(\boldsymbol{\beta}; t) = \left\{\sum_{k=1}^{K}\mathbf{s}_{jk}^{(2)}(\boldsymbol{\beta}; t)\right\} / \left\{\sum_{k=1}^{K}\mathbf{s}_{jk}^{(0)}(\beta; t)\right\}$.

(C) Let $\mathbf{s}_{jk}^{(d)}$, $d = 0, 1, 2$, $\mathcal{B}$ and $\mathbf{e}_j$ be as in Condition (B) and define $\mathbf{v}_j = \left\{\sum_{k=1}^{K}\mathbf{s}_{jk}^{(2)}(\boldsymbol{\beta}, t)\right\} / \left\{\sum_{k=1}^{K}\mathbf{s}_{jk}^{(0)}(\boldsymbol{\beta}, t)\right\} - \mathbf{e}_j(\boldsymbol{\beta}; t)^{\otimes 2}$. Then for all $\boldsymbol{\beta} \in \mathcal{B}$, $t \in [0, \tau]$, $j = 1, \ldots, J$ and $k = 1, \ldots, K$: $\mathbf{s}_{jk}^{(1)}(\boldsymbol{\beta}; t) = \partial\mathbf{s}_{jk}(\boldsymbol{\beta}; t)/\partial\boldsymbol{\beta}$ and $\mathbf{s}_{jk}^{(2)}(\boldsymbol{\beta}; t) = \partial\mathbf{s}_{jk}^{(1)}(\boldsymbol{\beta}; t)/\partial\boldsymbol{\beta}$. Assume $\mathbf{s}_{jk}^{(0)}(\boldsymbol{\beta}; t)$ is bounded away from 0 on $\mathcal{B} \times [0, \tau]$, and the matrix $\sum_{j=1}^{J}\int_0^\tau \mathbf{v}_j(\boldsymbol{\beta}_0; t)\sum_{k=1}^{K}\mathbf{s}_{jk}^{(0)}(\boldsymbol{\beta}_0; t)h_{0j}(t)\,dt$ is positive definite.

(D) In probability

$$\frac{1}{n}\sum_{i=1}^{n} E\{\|\mathbf{D}_i\|^2 I(\|\mathbf{D}_i\| > \varepsilon n^{1/2})\} \to 0.$$

These conditions are adapted from Clegg, Cai and Sen (1999) and guarantee the asymptotic normality of the pseudo-partial likelihood estimator, the maximizer of $\ell(\boldsymbol{\beta})$ defined in (5). Under these conditions, there exists a sequence $\boldsymbol{\beta}_n \to \boldsymbol{\beta}_0$ as $n \to \infty$.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, **19**, 716-723.

Antoniadis, A. and Fan, J. (2001). Regularization of wavelets approximations (with discussions), *J. Amer. Statist. Assoc.*, **96**, 939-967.

Badgonavičius, V. and Nikulin, M. (2002). *Accelerated Life Models: Modeling and Statistical Analysis*. Chapman and Hall, New York.

Bickel, P.J. (1975). One-step Huber estimates in linear models. *Jour. Ameri. Statist. Assoc.*, **70**, 428-433.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, **37**, 373-384.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection, *Ann. Statist.*, **24**, 2350-2383.

Cai, J. (1999). Hypothesis testing of hazard ratio parameters in marginal models for multivariate failure time data. *Lifetime Data Analysis*, **5**, 39-53.

Cai, J., Fan, J., Li, R. and Zhou, H. (2003). Variable Selection for Multivariate Failure Time Data. Manuscript.

Cai, J. and Prentice, R. L. (1995). Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika*, **82**, 151-164.

Cai, J. and Prentice, R. L. (1997). Regression estimation using multivariate time data and a common baseline hazard function model. *Lifetime Data Analysis*, **3**, 197-213.

Clegg, L. X. , Cai, J. and Sen, P. K. (1999). A marginal mixed baseline hazards model for multivariate failure time data. *Biometrics*, **55**, 805-812.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Jour. Roy. Statist. Soc. Ser. B*, **34**, 187-220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**, 269-276.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, **31**, 377-403.

Dawber, T. R. (1980). *The Framingham Study, The Epidemiology of Atherosclerotic Disease*, Cambridge, MA, Harvard University Press.

Donoho, D.L. and Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425-455.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with discussions), *Ann. Statist.*. **32**, 409-499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96**, 1348-1360.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, **30**, 74-99.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics*, **32**, 928-961.

Faraggi, D. (1998). Bayesian variable selection method for censored survival data. *Biometrics*, **54**, 1475-1485.

Faraggi, D. and Simon, R. (1997). Large sample Bayesian inference on the parameters of the proportional hazard models. *Statistics in Medicine*, **16**, 2573-2585.

Frank, I.E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109-148.

Giudici, P. Mezzetti, M. and Muliere, P. (2003). Mixtures of products of Dirichlet processes for variable selection in survival analysis. *Journal of Statistical Planning and Inference*, **111**, 101-115.

Ibrahim, J. G. and Chen, M. H. (2000). Power prior distributions for regression models. *Statistical Sciences*, **15**, 46-60.

Ibrahim, J. G., Chen, M. H. and MacEachern, S. N. (1999). Bayesian variable selection for proportional hazards models. *Canadian Journal of Statistics*, **27** 701-717.

Ibrahim, J. G., Chen, M. H. and Sinha, D. (2001). Bayesian variable selection for proportional hazards model. *Statistica Sinica*, **11**, 419-443.

Kim, S. W. and Sun, D. (2000). Intrinsic priors for model selection using an encompassing model with applications to censored failure time data. *Lifetime Data Analysis*, **6**, 251-269.

Lee, E. W., Wei, L. J. and Amato, D. A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, J. P. Klein and P. Goel, eds, Boston: Kluwer Academic Publishers, 237-248.

Liang, K.-Y., Self, S. G. and Chang, Y.-C. (1993). Modelling marginal hazards in multivariate failure time data. *J. Royal Statist. Soc., Ser. B*, **55**, 441-453.

Lin, D. Y. (1994), Cox regression analysis of multivariate failure time data: The marginal approach. *Statist. in Med.*, **13**, 2233-2247.

Lindley, D. V. (1968). The choice of variables in multiple regression (with discussion). *Jour. Roy. Statist. Soc., B*, **30**, 31-66.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics*, **15**, 661-675.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd ed. Chapman and Hall, London.

Miller, A.J. (2002). *Subset Selection in Regression*. 2nd Edition, Chapman and Hall, London.

Nielsen, G. G., Gill, R. D., Andersen, P. K., and Sø rensen, T. I. A. A. (1992). A counting process approach to maximum likelihood estimator in frailty models. *Scandin. J. Statist.*, **19**, 25-43.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461-464.

Sinha, D. Chen, M.H. and Ghosh, S. K. (1999). Bayesian analysis and model selection for interval censored survival data. *Biometrics*, **55**, 585-590.

Spiekerman, C. F. and Lin, D. Y. (1998). Marginal regression models for multivariate failure time data, *Jour. Amer. Statist. Assoc.*, **93**, 1164-1175.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *J. Royal Statist. Soc., Ser. B*, **58**, 267-288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model, *Statistics in Medicine*, **16**, 385-395.

Volinsky, C. T. and Raftery, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics*, **56**, 256-262.

Wei, L. J., Lin, D. Y. and Weisseld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J. Amer. Statist. Assoc.*, **84**, 1065-1073.

Index